DOCUMENT RESUME

ED 087 066 CS 500 578

AUTHOR	Su, Stanley Y. W.; Moore, Robert L.
TITLE	Discourse Synthesis, Analysis and Their Application to CAI (Computer Assisted Instruction).
INSTITUTION	Florida Univ., Gainesville. Communication Sciences Lab.
PUB DATE	Mar 72
NOTE	21p.; Communication Sciences Laboratory Quarterly Report, Volume 10, No. 1
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Computer Assisted Instruction; Content Analysis; *Discourse Analysis; Expository Writing; Higher Education; *Journalism; *Language Research; Linguistic Patterns; *Paragraph Composition; Persuasive Discourse; Semantics; Sentences; Writing Skills

ABSTRACT

This paper deals with the computer's production and recognition of sentences in a connected discourse and its application to computer assisted instruction. Studies of textual properties in real discourses have been carried out at the paragraph level. The theoretical concepts of representing paragraph content in terms of (1) the factual data expressed by the grammatical units in a paragraph, (2) the development types and their structural relationships and (3) the cohesive principles used in the sentences of the paragraph is presented. The theoretical investigation of discourse properties is aided by a paragraph generation system constructed to synthesize paragraphs and to test out the linguistic assumptions made in the study. The theoretical concept and the computational tool constructed are used in the development of an integrated computer-assisted system designed to synthesize and analyze news stories at the paragraph level and to provide preprogrammed critiques to students learning journalistic writing. (Author)

COMMUNICATION SCIENCES LABORATORY

QUARTERLY REPORT

DEPARTMENT OF SPEECH
UNIVERSITY OF FLORIDA

1972

VOL. 10 NO. 1
MARCH

Preface

The purposes of these progress reports are:

1. To provide other laboratory investigators and professional workers in the field with up-to-date information about our research activities and results,

2. To serve as documentation of our research activities for agencies which provide us with support, and

3. To provide somewhat formal reporting of research activity for our own faculty and students in order to exchange information and encourage collaborative efforts.

These reports are not intended to take the place of publications in recognized journals. Some contributions may be so published subsequently, but others may not. Many of these reports are of projects currently in process or with preliminary or partial findings being already available.

In view of these purposes and the nature of the contributions, no editorial review is exercised. Consequently, these reports should not be quoted unless specific permission to do so is granted by the author.

Inquiries concerning these reports should be addressed to the editor, Robert J. Scholes.

DISCOURSE SYNTHESIS, ANALYSIS AND THEIR APPLICATION TO CAI

Stanley Y.W. Su and Robert L. Moore

Abstract

This paper deals with the computer's production and recognition of sentences in a connected discourse and its application to computer assisted instruction. Studies of textual properties in real discourses have been carried out at the paragraph level. The theoretical concepts of representing paragraph content in terms of (1) the factual data expressed by the grammatical units in a paragraph, (2) the development types and their structural relationships and (3) the cohesive principles used in the sentences of the paragraph is presented. The theoretical investigation of discourse properties is aided by a paragraph generation system constructed to synthesize paragraphs and to test out the linguistic assumptions made in the study. The theoretical concept and the computational tool constructed are used in the development of an integrated computer-assisted system designed to synthesize and analyze news stories at the paragraph level and to provide preprogrammed critiques to students learning journalistic writing.

Introduction

Studies of units of discourse larger than the sentence done by scholars in languages and linguistics is represented by such diverse fields as discourse analysis (Harris, 1963), beyond-the-sentence analysis (cf. Hendricks, 1967), the linguistic study of literary texts (Bailey, 1968, 53-76), stylistic analysis (cf. Sedelow, 1968) and text analysis (Oomen, 1971). With the exception of Harris' work, many reported studies have dealt with philosophical discussions of discourse analysis or statistical accounts of word usage in texts rather than with the study of the formal properties of discourse. By formal properties, we mean the specific rules which account for the language user's production and recognition of sentences in a connected discourse and which can be processed by a digital computer.

Most modern linguists do not regard text as a proper unit of linguistic concern. On the one hand, text meaning has been regarded by some linguists as a problem that falls in the domain of language "performance" which, according to Chomsky (1965), is not the immediate concern of linguistic study. On the other hand, text has been considered by others to be a linguistic unit that can be derived from sentences by using proper connectors such as "and," "but," "however," etc. Therefore, these linguists feel sentences rather than text should be the proper unit of linguistic concern. Recent works opposing these points of view can be found in Daněs 1970 and Oomen 1971.

Computer science has great potential applicability to linguistic study, and the syntactic and semantic problems beyond the sentence boundary have been examined by computer scientists. The techniques of both synthesis and analysis have been used in many computer systems to deal with problems related to text. Works reported by Klein (1965a, 1965b), Weizenbaum (1966, 1967), Schank (1968), Su and Harper (1969), Friedman (1969), Vigor, Urguhart and Wilkinson (1969), Woolley (1969) and Su (1971b) are concerned with the automatic generation of coherent sentences and texts. A few of the existing question-answering systems (cf. Simmons, 1970) and Wilks' semantic analysis of English paragraphs (1968, 1969) are concerned with the automatic recognition of the information content of sentences in a connected discourse. The present work is a natural outgrowth of these previous studies.

The linguistic properties of connected discourse is still poorly understood despite past efforts by both linguists and computer scientists. The lack of knowledge concerning the formal properties of language elements beyond sentence boundaries has hindered progress in many areas of study in computer science.

In most computer-assisted instruction (CAI) application for example, the student is expected to input his answers to the system according to fairly rigid system- and author-defined requirements (Spolsky, 1966; Zinn, 1967; Silberman, 1968; Paulus, et al., 1969 and U.S. Department of Commerce 1970). Unanticipated student responses will cause the system to supply the student with more information and ask him to try again.

This kind of analysis and feedback is clearly insufficient for instruction requiring textual input from the student (Spolsky, 1966; Paulus, et al., 1969; Moore, 1972). In language or writing instruction, for example, part or all of the student's answer may be hierarchically or synonymously related to the system-stored "correct" answer, but may be unrecognizable to the computer and thus judged incorrect. Even a paraphrasing or simple reordering of the student's answer could be interpreted as incorrect if it did not match the stored answer. Systems such as these, by failing to analyze the structure and semantic content of the student's answer, are losing much valuable information that would add precision to the instructional process.

In the areas of content analysis and automatic abstracting, a common method used is to construct a word-frequency table and to statistically select key-words that best describe a document's contents (IBM Data Processing Techniques, n.d.; Meadows, 1970, pp. 114-20). A more sophisticated approach introduced by the French Syntol group (Cros, et al, 1964) is to select from a text formal entities known as "Syntol words" and to connect them in semantic diagrams with one of four specific relations to represent the contents of the text. However, the analysis is mainly at the sentence level, and it therefore loses the semantic relationships existing among concepts which cross the sentence boundary. It is also noted that the original construction of such diagrams is substantially an unsolved problem, since the depth of analysis must be greatly increased if additional relational types are to be identified (Salton, 1968,p.230).

Two obvious problems associated with most existing information storage and retrieval (IS&R) systems are (1) the restrictions placed upon the index language, i.e., rigid index terms (keywords) and simple phrases or sentences, and (2) the restrictions placed upon the query languages (Meadows, 1970, p. 180). The existing natural-language information retrieval systems attempt to remove these restrictions by analyzing text and natural language queries at the sentence level. However, many problems of ambiguity can not be resolved at the sentence level. Moreover, the failure to recognize the development and cohesive properties in a document text and in queries results in poor retrieval precision. The following query will help illustrate this point:

"List the titles of all documents written since 1965 that show how the learning theory used in programmed instruction is applied to computer-assisted instruction."

One of the main components of this request is "is applied to." The user in this case does not want documents that merely discuss both programmed instruction and CAI; he wants those which show how learning theory in the former is applied to the latter. Only in a system which recognizes the development properties, "action" and "means of performing the action," existing in a text will the needed precision be expressible and recognizable.

In previous research (Su and Harper, 1969, Su 1971b), we have studied text at the paragraph level and have successfully developed a computational tool, a paragraph generation system, for testing the linguistic assumptions made concerning the properties and relationships among grammatical units in paragraphs. Many sentence patterns and cohesive principles have been observed in the analysis of physics texts, and have been tested and evaluated in the generation system designed and implemented for the study.

This research has been extended to the investigation of linguistic properties in journalistic writing. In our present study, students' and instructors' short news stories written for courses on journalistic writing are analyzed to determine the structural and semantic relationships among basic grammatical units in the continuous texts. The relationships observed are formalized in terms of writing "rules" and are tested in the generation system.

The theoretical concept and the knowledge gained from experiments with the paragraph generation system are applied to the investigation of problems related to recognition of paragraph content. A discourse analysis system is being developed to test the analysis rules constructed in the study. It is being incorporated into the existing paragraph generation system and an existing computer-assisted instruction component to form a powerful computer-assisted instruction system. This integrated system is designed to teach students the basic concepts of journalistic writing.

This paper describes the theoretical concept of representing paragraph content by the attributes of development and cohesion, and outlines the paragraph generation system developed for the study. The research in progress on the integrated computer-assisted instruction system is described in the last section.

A Model of Paragraph Production and Recognition

It has been our contention that the semantic content of a paragraph is represented by a set of basic concepts characterized by the attributes of "development" and "cohesion" (Su, 1971b). In our model, we distinguish among three types of paragraph constituents. Type I constituents are grammatical units, such as sentences, noun phrases, prepositional phrases, etc., which express concepts or factual data in a paragraph. These basic concepts are interrelated through some abstract relations which represent textual development or progression through the paragraph in the form of spatial, temporal and logical movement. We shall call these abstract relations the development types or Type II constituents. Some intersentence connections are: (1) progression from the general to the specific, (2) time progressions, for example from past to present, from present to future, (3) location changes from one place to another, (4) cause and effect, (5) action to means of performing the action, etc.

These relations are introduced by some specific syntactic patterns as well as semantic properites of words in single sentences, adjacent sentences or non-adjacent sentences in the paragraph. The following examples illustrate some of the patterns and word properties which introduce respectively the development types listed above:

(1) John investigated the case carefully. He questioned (more specific action than investigated) Tom regarding the fire.

(2) In the work reproted by Smith, ... The present paper describes . . . .

(3) He moved from Chicago to New York.

(4) They trained very hard for the game. As a result, they won the division title.

(5) He stopped the leak by inserting a stick into the hole.

A development type can be described by many single-sentence frames or sequences of sentence frames. A sentence frame is a sentence skeleton, and it specifies partial syntactic structure and/or semantic properties of a sentence which, together with those of ther sentences in a paragraph, bring about the development type. For example, causation is a development type marking the relation of causation between two concepts A and B, which are called the arguments of the development type. It may be described by a set of basic single-sentence frames and/or sequences of sentence frames. Some examples of single-sentence causation frames are "B because of A," "A accounts for B", "A leads to B," "B is due to A," etc. Some examples of sequences of sentence frames are "A. Therefore, B", "B. As a result of A," "A. Thus, B," etc.

In the above examples, the sentence frames specify the syntactic and semantic constraints in which the arguments A and B are presented. Thus, arguments A and B may be expressed in the forms of phrases, sentences or sequence of sentences depending upon the particular frame in which they occur. Since there can be many different single-sentence frames and sequences of sentence frames associated with a development type, a number of paragraphs can be produced which have the same semantic content-- the same basic concepts but with different syntanctic forms.

In our inplementation, single-sentence frames and sequences of sentence frames associated with a development type are stored in the computer as dependency structures. The nodes on a dependency structure specify restructions in terms of syntactic categories, semantic classes or specific words. Therefore, a sentence or sequence of sentences of any length that satisfies the restructions would represent the development type. The dependency structures will be referred to in the later sections as "restriction patterns" and are used as one source of constraints under which paragraphs are produced. They also form the pattern dictionary for the analysis system under implementation. Since development types are described in terms of both syntactic patterns and semantic properties, they can be regarded as both syntactic and semantic constituents of the paragraph. Patterns which are used to introduce development types can be considered as discourse dependent, whereas development types are discourse independent and are textual properties that can occur in all discourses.

More than one development type can be employed in a paragraph to specify the high level conceptual relationships among the set of basic concepts in the paragraph, and they can be related to one another in a complex structure. The structural relationship among development types is called the development structure of the paragraph and constitutes a part of syntactic description of the paragraph. We visualize that the structure of the paragraph can be represented by the following production rules:

(1) Paragraph → Development (Development)

(2) Development → Causation ⟨Argument (Argument)⟩ |

Time ⟨Argument (,Argument)⟩ |

Means ⟨Argument (,Argument)⟩ |

.

.

.

(3) Argument → Paragraph | Variable

(4) Variable → Grammatical Units

The rules shown above indicate that a paragraph is composed of one or optionally (parenthesis) more than one development which can be rewritten as one of the alternative development types each of which is indicated by a name followed by a list of arguments enclosed in meta-symbols '<' and '>'. The vertical bars indicate the alternatives. The listing of alternative development types in Rule (2) above can be considered as non-terminal symbols each of which can be rewritten as a set of single-sentence frames or a set of sequences of sentence frames. An argument of a development type can itself be a paragraph or a variable. Thus paragraph developments can be nested. A variable can be rewritten as any grammatical unit which satisfies the constraints of the sentence frames associated with the development types.

The rules proposed above account for the imbeded structural relationships among tokens of the same development type (Causation) as expressed in potentially infinite number of derived sentence frames such as "C is due to D," "B leads to C is due to D," "A because of B leads to C is due to D," etc. They also account for the hierarchical relationships among different development types as shown in the framework of paragraph and development structure (Figure 1).

"Smith's early research has been proven wrong by recent studies. . . . We therfore conclude that . . . . "

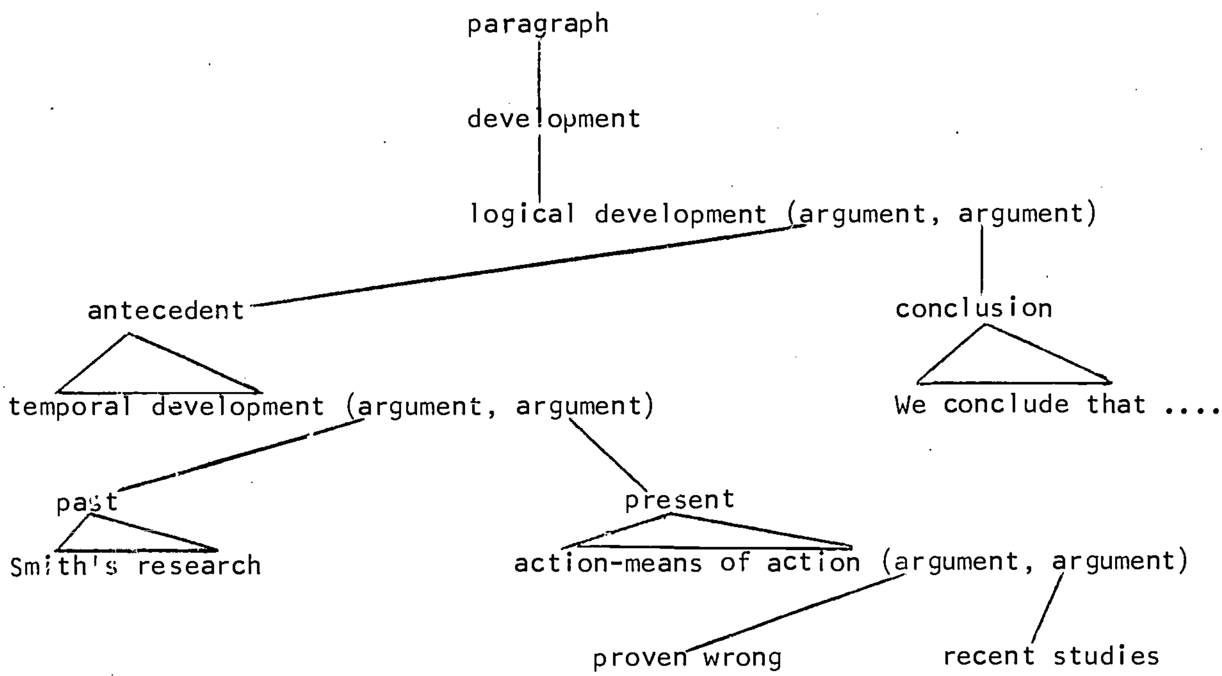

Figure 1. A development structure

The third constituent (Type III) of a paragraph is the group of cohesive principles applied in writing to achieve continuity and relatedness among the words which form the Type I constituents. Some of the principles of cohesion commonly employed in sentence and intersentence connection are (1) word repetition, (2) use of anaphoric words, (3) use of hyperonyms, synonyms and antonyms, (4) use of ellipses and (5) use of words that are partially semantically related where relatedness can be defined in terms of the hierarchical relationships of their semantic features. The following examples illustrate the use of common cohesive principles.

(1) Two armed gunmen, today, held up a .....
The gunmen fled on foot, leaving no trace ....(word repetition)

(2) John bought a new bicycle. He rode it to school today.
(anaphoric reference)

(3) Mary wore a rose today. The flower was given to her by her boy-friend. (hyperonym)
She finally met the bachelor. But the unmarried man disliked her. (synonyms)
Now the younger one is happy but the older one is sad. (antonyms)

(4) He has a new car, a good stereo set and a beautiful house. All these were given to him by his father. (ellipsis)

(5) Brody threw three touchdown passes today. The game was played before a shivering crowd of 55,000. (words in the same semantic field)

There are other cohesive devices commonly used in writing that may not be as apparent as the examples shown above. These are the literary methods used by writers to relate possibly abstract notions to the readers' knowledge of the world. Such devices include specific details and examples, metaphor, analogy, simile, imagery, personification, animation, anthropomorphism, synecdoche, and so on.

Cohesive principles account for the semantic relatedness among concepts expressed in adjacent or near-adjacent sentences in a text. Some of the principles listed above such as pronominalizations and lexical repetitions are traditionally considered to be part of syntax and can be introduced by syntagmatic substitutions. Other principles are generally considered to be semantics. A computer system which recognizes these principles in a text may use them as a basis for resolving word or sentence ambiguities by selecting the meaning which conforms with the principles.

The basic theoretical assumption of our computational model for paragraph production and recognition is as follows: A speaker has initially some basic concepts which he intends to express. These concepts can be expressed in phrases or sentences which can be generated by a generative grammar such as a phrase structure grammar or a dependency grammar. The relationships among the basic concepts can be described by what we have called development types and principles of cohesion. In producing a paragraph, the speaker first chooses a familiar sentence frame or a sequence of sentence frames associated with each development type he intends to express. Phrases and sentences which express the basic concepts are then chosen to fit the abstract structural and semantic framework selected for the development types. In selecting the words to make-up the phrases and sentences, the principles of cohesion are followed to achieve semantic continuity and lexical relatedness of the resulting paragraph.

Recognition of a paragraph content is considered as a process of identifying the proper semantic interpretation of the set of grammatical units (Type I constituents) and their relations defined in terms of development types (Type

II) and cohesive principles (Type III).

A language user has a set of sentence frames or sequence of sentence frames that are familiar to him. In reading a paragraph or listening to a sequence of utterances, when the words or phrases he encounters have syntactic and semantic properties which match part of the pattern he knows, he can then predict the occurrence of the rest of the patterns and knows the development types that are introduced. For example, when one hears or reads the phrase "on the one hand," he can with certainty predict the use of the phrase "on the other hand," and he knows that the abstract concept of contrast is being introduced to relate some facts. The recognition of development types and their structural relationships allows a language user to follow the progression of events.

The cohesive principles are schemes which the language user uses to introduce semantic relationships between adjacent or near-adjacent sentences. The recognition of these principles in utterances or writings allows one to resolve ambiguities existing at the sentence and the beyond-the-sentence level and to relate factual data to proper referents.

What actually takes place in the language user's mind when he reads a paragraph or listens to somebody talking is more than one can say. However, we have used the simple model presented as a first approximation to develop a computer system for testing our assumption. It is hoped that through the study of text in a real discourse and the use of computer as a tool to test some of the textual properties observed, we can obtain better understanding on this difficult, yet very important, subject on the properties of discourse.

The Design and Implementation of A Paragraph Generation System

One way to test the production aspect of our model would be to build a computer system which is supplied with a large lexicon, a grammar for producing sentences and a set of sentence frames for describing development types to produce a potentially infinite number of paragraphs using the proposed rules shown in the preceeding section. This set would contain paragraphs having development and cohesion characteriestics which have occurred in existing texts as well as those which may occur in any as yet unwritten text. However, this nearly uncontrolled production is expensive and the evaluation of the produced paragraphs can be very difficult. An alternative method, which is more economical and in our opinion more interesting, is to supply the system, in addition to a grammar, with words and word behavior data compiled from a real discourse and to implement the development types and principles of cohesion (to be evaluated by the system) to control the paragraph production. The sentence frames and sequence of sentence frames used in the real discourse to introduce various development types are supplied to the system. The produced paragraphs can then be checked against the paragraphs in the real discourse to determine the validity of development types and principles of cohesion implemented. To further control the paragraph production, the system can be given a paragraph theme containing the development types which are to appear in the generated paragraphs.

The evaluation of the output from this controlled production leads to the modification of the sentence frames associated with the development types and of the principles of cohesion implemented in the system. These modifications, will serve as the basis for further generations. This section outlines the design and the implementation of a paragraph generation system which carries out these ideas. For more detailed description, the reader is referred to the previous publications (Su and Harper, 1969 and Su, 1971b).

The generation system consists of two major components: a restriction pattern constructor and a paragraph generator. Input to the pattern constructor is the user's specification of the theme (of a paragraph) which he wants the generated paragraph to contain. The theme of a paragraph is represented by one or more than one type of paragraph development and some specific lexical items. For example, compare A with B can be a theme of a paragraph where comparison is a type of paragraph development and A and B are specific lexical items. Associated with each type of paragraph development are a number of sequences of restriction patterns provided to the system. Each sequence specifies the global syntactic and semantic information of a string of sentences which constitutes the development type. The function of the pattern constructor is to select some sequences of restriction patterns (one for each development type specified by the user) and to combine them to form a single sequence of restriction patterns which possesses all the global syntactic and semantic information of the selected sequences. The resulting sequence is used as input to the paragraph generator, and it controls the generation process to produce the specified development types in the output paragraph.

The paragraph generator generates each sentence of a paragraph under the direction of the following parameters:

(1) A restriction pattern which specifies, in the form of a tree structure, the partial syntactic and semantic properties of the sentence to be generated. The nodes on the tree structure are restrictions specified in terms of semantic markers, specific lexical items or syntactical features. The generator generates a sentence under the constraints of the specified restrictions which bring about the development of a paragraph.

(2) Parameters used for achieving semantic cohesion in the generated paragraph: the probability of co-occurrence of word classes; semantic classification of lexical items based on word distribution, word family, synonyms, and words in the same semantic field; implemented criteria such as word repetition, use of anaphoric words, substitution with hyperonyms and synonyms and the like. The generator increases or decreases the weight of each allowable constituent on the basis of these parameters, and it selects the word which has the highest reweighting value. Random selection is applied only when there is more than one word with the same reweighting value.

The system uses a dependency grammar and operates with English language materials. It is not oriented towards or restricted to a particular language or set of language data. Programs are written in PL/1 and have been run on an IBM 360 Model 65. A series of experiments with the generation system have been conducted and some experimental results and considerations were presented in earlier papers.

An Integrated Computer-Assisted Instruction System

Our research in progress is the continuation and extension of work described in the preceding sections. It focuses on the study of discourse structures and semantic properties in journalistic writing and on the implementation of a paragraph analysis system to be incorporated in an existing CAI system.

In this section, we shall show how the theoretical study of paragraph properties and its application to a real world problem in education can be achieved in an integrated computer system consisting of a paragraph generation component, a discourse analysis component and a CAI component. Figure 2 shows the components of the system. Their functions are described together with the description of our research tasks.

In beginning journalistic writing courses, students are given or are led to discover certain "rules" (Label 1 in Figure 2) and stylistic patterns which will govern the manner in which they write news stories. They are asked, for example, to consider reader interest, news value, organization, variety, kinds of leads and so on. Based upon these journalistic rules and considerations, we analyze a set of short news stories written by students and instructors in courses on journalistic writing to determine the development types and cohesive principles followed by the writers. The inter-sentence patterns and semantic relationships among words and phrases which introduce the development types and cohesion principles are determined.

The sentence patterns are stored as dependency structures (Label 2) in a pattern dictionary. The nodes of these patterns specify syntactic or semantic constraints in terms of semantic classes (or markers), syntactic features or specific lexical items. A paragraph which matches a particular structure and satisfies its constraints will thus indicate the development types with which the patterns are associated. Since each node of the patterns can be specified in terms of a multiplicity of semantic markers and/or syntactic features, a single pattern may match with many paragraphs which are semantically equivalent with respect to paragraph development but quite different syntactically.

We presently have a 2000 word dictionary and rapid search algorithm based upon frequency of occurrence data of standard American English (Kucera and Francis, 1967). The computer representation of rules, together with the lexicon, are given to the existing paragraph generation system (Label 4) to produce paragraphs for evaluation by the researchers. If deficiencies are observed in the output paragraph, the formal representations of rules are modified and used as basis for further generation until the formalized rules will lead to the production of desirable paragraphs. During this feedback process, the lexical entries are also modified to ensure that proper syntactic and semantic features have been assigned to them. The analyzer (Label 5), under development, analyzes text at the paragraph level and produces analysis results for each paragraph of the input text. The structured development types, the cohesive principles and the semantic relationships of the content words in a paragraph represent the contents of the paragraph.

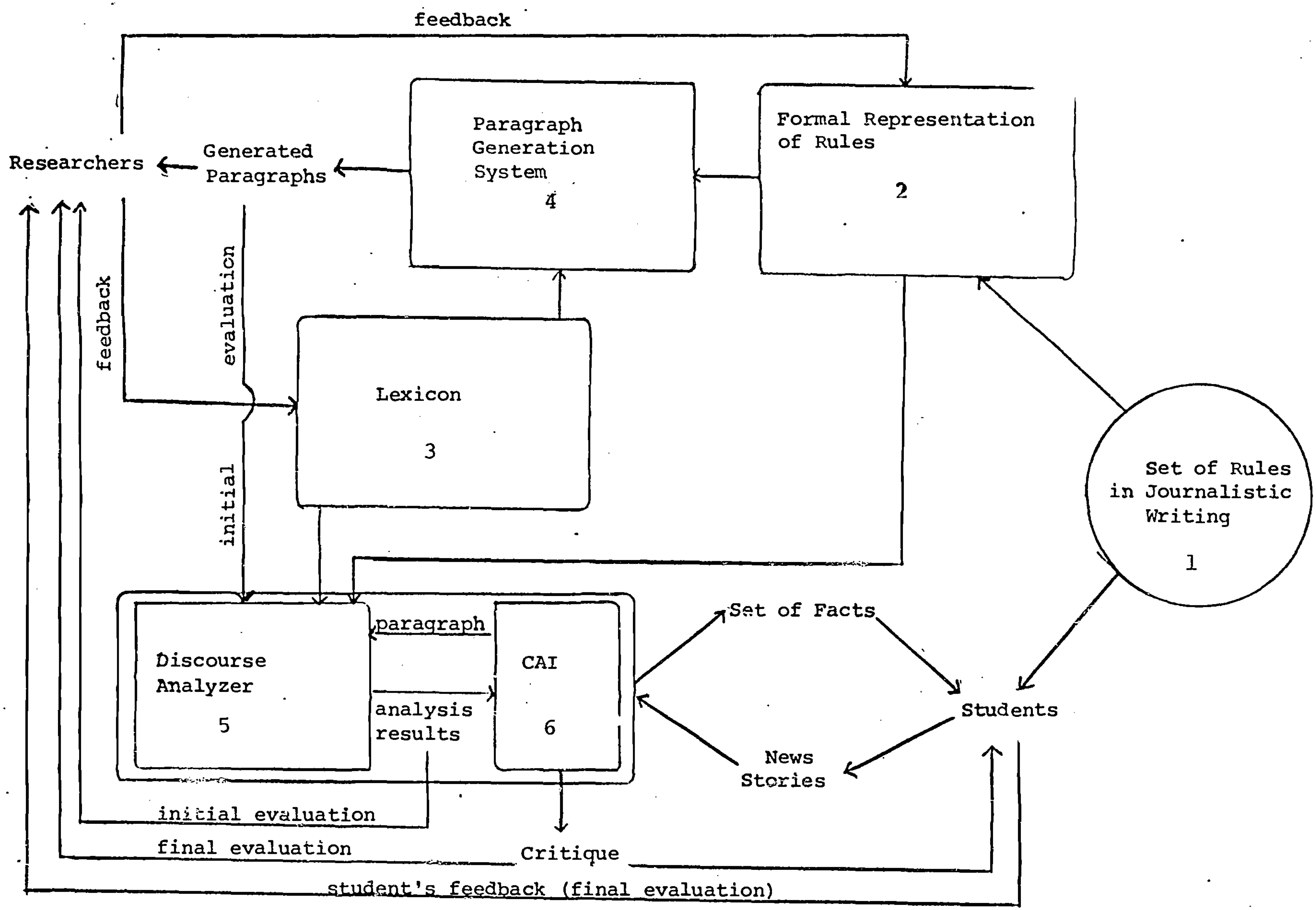

FIGURE 2. System Configuration

To determine the development types and their structures in a paragraph, the analysis system follows this procedure:

1. A dictionary look-up procedure retrieves the syntactic and semantic information for the words and phrases in the input. Phrases such as "on the one hand," "on the other hand," "it is known that," "it is necessary that," etc. are examples of development indicators which indicate that the text contains certain development types. Such phrases are treated as single lexical items.

2. The sentences of an input text are then analyzed using a dependency grammar to produce their surface structures.

3. The pattern dictionary established for a particular discourse, journalistic writing in this case, is then used in a matching procedure for determining the development types and their structural relationships within the paragraphs of the input text.

4. The result of this matching procedure is a development structure in which a high degree of development complexity in paragraphs can be represented. The development structure is represented as a hierarchical tree on which the nodes represent the development types and the leaves represent the phrases and sentences which are the arguments of the development types.

5. The phrases and sentences which express the basic concepts of the paragraph are analyzed to determine the semantic relationships among their constituents. These relationships are determined on the basis of anaphoric relations, word repetition, synonyms, antonyms, words of same semantic classes or classes with intersecting h rarchies and other cohesive principles. The use of logical properties such as set-membership, set inclusion, transitivity, symmetry and other relations used in some existing question-answering systems (cf. Simmons 1970) is contemplated.

The development structure of a paragraph and the phrases and sentences expressing its basic concepts are entered in a data base containing other paragraphs of a text. The data structure used to represent the contents of a text is an associative structure in which items (basic concepts) and development types are linked associatively and the links are labelled with specific relation names. Retrieval algorithms and data manipulation procedures for such data structure are reported in Kay and Su (1970), and Su (1971a).

For initial evaluation of the analyzer, we feed the output of the generation system to the analyzer. If the discourse analyzer is working properly, it should recognize the inter-sentence patterns and semantic relationships among lexical items which are the formal representation of the rules used by the generation system to produce the input paragraphs. The deficiencies observed in the output of the analyzer will be used, through the researchers' feedback channels, to modify the lexicon and the formalization of rules.

An existing CAI system (label 6), using IBM's Coursewriter III available at the University of Florida, is being incorporated into the system to serve as interface between the discourse analyzer and the students. The CAI component is capable of displaying preprogrammed facts, accepting news stories, and generating preprogrammed critiques. The analyzer and the CAI component will be used to critically analyze journalistic writing in the following manner: A student who has been exposed to the general rules of journalistic writing is given a set of facts related to a real or hypothetical news event. From this set of facts he is expected to write a news story and input it through an on-line terminal to the system. The discourse analyzer is called to critically analyze each paragraph in terms of cohesion, development, transition and logical ordering. Nested within these measures are judgments of such elements as repetition and redundancy and, later in our research, literary effects such as metaphor. The system will, by using semantic class information, recognize the use of clichés, unnecessary adverbial clauses, misplaced modifiers, etc. The system will also analyze "nonspecific" writing (the use, for example, of weak nouns and verbs such as "dog" and "ran" in place of " the small terrier" and " scurried," respectively).

The analysis results are printed to the researchers for evaluation and are passed to the CAI component which in turn generates a critique to the student on the basis of analysis results.

The system is designed to operate on an interactive, real-time basis so that the student can then make any necessary changes and corrections to his story and resubmit it immediately.

The student's experience with the system, the critique and the analysis results are the information sources for the final evaluation of the whole system. Through such experiments, the researchers can tune the system through the feedback channels. The process can then be repeated.

Journalistic writing is well suited to CAI of this type. Journalism students are trained to write news stories in an "inverted pyramid" style--that is, the most important facts are presented first followed by material in descending order of importance. A CAI system based upon paragraph description could be readily programmed to recognize this style of writing since we could, by defining certain restriction patterns, specify to the system the kind of writing structure we are expecting. Such a well-structured application of the analyzer will form a solid theoretical foundation for later CAI applications.

The generation and analyzer systems have applicability to discourse analysis in areas other than CAI. Computer-based content analysis, information storage and retrieval, language translation and question-answering are all presently hampered, we feel, by the inability to detect beyond-the-sentence development and cohesion. The methodology we employ could be used, for example, to provide abstracts which would have the same high-level semantic framework as the original document and thus give the user a clearer idea of its contents than is possible with existing systems.

As our research continues, we will gain insight into discourse analysis techniques and we will be able to provide a more complete description of formal properties of paragraphs than is presently available.

Bibliography

Abelson, R.P., and Reich, C.M. Implicational Molecules: A Method for Extracting Meaning from Input Sentences. Proc. Int. Jt. Conf. Art. Intel., Washington, D.C., 1969, pp. 641-647.

Arsent'eva, N.G. O Dvux Sposobax Porozdenija Predlozenij Russkogo Jazyka. Problemy kibernetiki, No. 14, 1965, pp. 189-218. ("Two Means of Generating Sentences in Russian." Translation in JPRS 32, 605, Oct. 28, 1965, pp. 272-314.)

Bailey, R., and Burton, S. English Stylistics: A Bibliography. M.I.T. Press, 1968, pp. 53-76.

Bishop, Robert L. A Basic Course in Writing and Computer Analysis of Natural Language Text for Style and Content in the Context of Instruction in Writing. Available from Dr. Bishop, Department of Journalism, University of Michigan Ann Arbor, Michigan.

Bishop, Robert L. Learning to Write--From a Computer. Quill, May, 1971. pp. 22-23.

Bobrow, D.G., A Question-Answering System for High School Algebra Word Problem. Proc. AFIPS, 1964 Fall Joint Computer Conf., Vol. 26, Pt. 1, Spartan Books, New York, pp. 591-614.

Chomsky, N. Aspects of the Theory of Syntax. Cambridge, Mass. 1965.

Chomsky, N. Deep Structure, Surface Structure and Semantic Interpretation. M.I.T. Report. 1969; also in R. Jacobson and S. Kawamoto (eds.) Studies in General and Oriental Linguistics Presented to Shiro Hattari. TEC Co., Tokyo, pp. 52-91.

Chomsky, N. Remarks on Nominalization, In R. Jacobs and P. Rosenhaum (eds.) Readings in English Transformational Grammar. Waltham, Mass. 1970.

Colby, K.M., Tesler, L., and Enea, H. Experiments with a Search Algorithm for the Data Base of Human Belief Structure. Proc. Int. Jt. Conf. Art. Intel., Washington, D.C., 1969, pp. 649-654.

Coles, L.S. An On-Line Question-Answering System with Natural Language and Pictorial Input. Proc. ACM 23rd Nat. Conf. 1968, Brandon Systems Press, Princeton, N.J., pp. 157-167.

Coulson, John E., CAI and Its Potential for Individualizing Instruction. Washington, D.C.: Academy for Educational Development, Inc. 1970.

Cross, R.C., Garden, J.C. and Levy, F. Syntol-Syntagmatic Organization Language. Gautheier-Villars, Parris, 1964.

Daneš, František. FSP and the Organization of the Text. Paper presented at the Conference on Functional Sentence Perspective at Marienbab, Czechoslovakia, 1970.

Dastert, B.H., and Thompson, F.B. How Features Resolve Syntactic Ambiguity. Proceedings of the Symposium on Information Storage and Retrieval. University of Maryland, 1971.

ENTELEK, Inc., CAI Information Exchange, Final Report, ENTELEK, Inc., Newbury Port, Mass. 1970;

Feldhusen, John H., and Lorton, Paul Jr., A Position Paper on CAI Research and Development. A Series Two Paper from ERIC at Stanford, 1970.

Friedman, Joyce, Directed Random Generation of Sentences. CS-80 AF-15, Stanford University Computer Science Department, October 1967; also in Communications of the ACM, 12, 1, January 1969, pp. 40-46.

Green, C.C. Application of Theorem Proving to Problem Solving. Proc. Int. Jt. Conf. Art. Intel., Washington, D.C., pp. 219-239.

Hansen, Duncan N. Current Research Development in CAI. Florida State University, Tallahassee, CAI Center, 1970.

Harper, K.E. A Study of the Combinatorial Properties of Russian Nouns. Mechanical Translation. August 1963, p. 36.

Harper, K.E. Measurement of Similarity Between Nouns. The RAND Corporation, RM-4532-PR, May 1965.

Harper, K.E. Syntactic and Semantic Problems in Automatic Sentence Generation. Second International Conference on Computational Linguistics, Grenoble, 1967.

Harris, Z.S. Discourse Analysis Reprints. Mouton and Co., The Hague, 1963.

Hays, D. Computational Linguistics. Elsevier, 1964.

Hendricks, W.O. On the Notion 'Beyond the Sentence'. Linguistics, 37, Dec. 1967, pp. 12-51.

IBM, Coursewriter III, Version 3, Author's Guide (5736-E11). 1971.

IBM, Data Processing Techniques, Keyword-In-Context (KWIC) Indexing, (GE20-8091-0), undated.

Kasher, A. Data Retreival by Computer: A Critical Survey. Hebrew U. of Jerusalem, Jan. 1966, Tech. Rep. No. 22 to Off. of Naval Res., Inf. Syst. Branch.

Katz, J.J. Recent Issues in Semantic Theory. Foundations of Language. III. 1967, pp. 124-194.

Katz, J.J. and Fodor, J. The Structure of a Semantic Theory. In J. Fodor and J.J. Katz (eds.) The Structure of Language. Englewood Cliffs, N.J., 1964.

Kay, M. and Su, Stanley Y.W. The MIND System: The Structure of the Semantic File. RM-6265/3 PR, The RAND Corporation, Santa Monica, California, 1970.

Kennam, E.L. Two Kinds of Presummposition in Natural Language. In C.J. Fillmore and D.T. Langendoen (eds.) Studies in Linguistic Semantics. Holt, Reinhart and Winston, Inc. 1971.

Kellogg, C.J. Burger, J., Diller, T. and Fogt, K. The Converse Natural Language Data Management System: Current Status and Plans. Proceedings of the Symposium on Information Storage and Retrieval. University of Maryland, 1971.

Klein, S. Automatic Paraphrasing in Essay Format. Mechanical Translation Vol. 8, Nos. 3 and 4, June and October, 1965.

Klein, S. Control of Style with a Generative Grammar. Language Vol. 41, No. 4, 1965.

Kochen, M. Automatic Question-Answering of English-Like Questions About Simple Diagrams. Journal of ACM, Vol. 16, No. 1, Jan. 1969, pp. 26-48.

Kucera, Henry and Francis, W. Nelson. Computational Analysis of Present-Day American English. Providence, Rhode Island: Brown University Press, 1967.

Lakoff, G. On Generative Semantics. To appear in D. Steinberg and L. Takobovits (eds.) An interdisciplinary Reader in Philosophy, Linguistics, Anthropology and Psychology. Cambridge University Press, N.Y. (In press).

Langendoen, D.T. and Savin, H.S. The Projection Problem for Presuppositions. In C.J. Fillmore and D.T. Langendoen (eds.) Studies in Linguistic Semantics. Holt, Rinehart and Winston, Inc. 1971.

Lekan, Helen A. (ed), Index to CAI. Milwaukee, Wisconsin, Milwaukee Instructional Media Lab., University of Wisconsin, 1970.

Lomkovskaja, M.V. Iscislenie, porozdajascee jadernye russkie predlozenija. Naucno-Texniceskaja Informacija, No. 7, 1965, pp. 35-41; No. 9, 1965, 37-40. ("Generation of Kernel Sentences in Russian." Translation in JPRS, 34, 159, 1966, pp. 104-113.)

McCawley, J.D. The Role of Semantics in Grammar. In E. Bach and R. Harms (eds.) Universals in Linguistic Theory. New York. 1968.

McDavid, Rauen I. Analysis of Natural Language in Wayne H. Holtzman Computer-Assisted Instruction, Testing and Guidance, New York: Harper and Row, 1970, pp. 222-227.

Meadows, Charles T. Man-Machine Communication. New York, Wiley, 1970.

Minker, J. and Sahle, J.D. Relational Data System Technology - A Survey and Critique. RADC Report, 1971.

Minsky, M. (ed.) Semantic Information Processing. MIT Press, Cambridge, 1969.

Moore, R.L. An Investigation of the Use of Computer-Assisted Instruction to Teach Elementary Statistics to Graduate Students. MAJC Thesis, Gainesville, Florida, College of Journalism and Communications, 1971.

Moore, Robert L. Computer-Managed Instruction: Toward Individualized Instruction. Communication Research Center, College of Journalism and Communications, University of Florida, Gainesville, Florida, 1972.

Morgan, J. On the Treatment of Presuppositions in Transformational Grammar. In B.J. Darden et al (eds.) Papers from the Fifth Regional Meeting of the Chicago Linguistics Society. Chicago. 1969.

Olney, J.C. and Londe, D.L. An Analysis of English Discourse Structure, With Particular Attention to Amaphoric Relationships. SP-2769, SDC, Santa Monica, Calif, 1967.

Oomen, Ursula. New Models and Methods in Text Analysis. In Richard J. O'Brien, S.J. (ed.) Monograph Series on Languages and Linguistics. Washington, D.C.: Georgetown University Press, 1971.

Parker, William Riley (ed.), The MLA Style Sheet. Revised Edition, New York: The Modern Language Association of America, 1968.

Paulus, Dieter, H. Some Applications of Natural Language Computing to Computer-Assisted Instruction. Contemporary Education: Computer-Assisted and Multi-Media Instructional Systems, Vol. XL, No. 5, April, 1969, pp. 280-285.

Quillian, M.R. The Teachable Language Comprehender. Comm. ACM, 12, 8, (Aug. 1969), pp. 459-476.

Sakai, T. and Nagao, M. Sentence Generation by Semantic Concordance. Presented at 1965 International Conference on Computational Linguistics, New York, 1965.

Salton, G. Automatic Information Organization and Retrieval. McGraw-Hill Book Company, New York, 1968.

Schank, R.C. Outline of a Conceptual Semantics for Generation of Coherent Discourse. TRACOR 68-462-V. 1968.

Scholes, R.J. On the Spoken Disambiguation of Superficially Ambiguous Sentences. Language and Speech. Vol. 14, Part 1, 1971. pp. 1-11.

Sedelow, S.Y. and Sedelow, W.A., Jr. Stylistic Analysis. In H. Borko (ed.) Automated Language Processing. Wiley, 1968, pp. 181-213.

Silberman, H.F. and Filep, R.T. Information Systems Applications in Education in Carlos A. Cuadra (ed.), Annual Review of Information Science and Technology American Society for Information Science, Vol. 3. Chicago: Encyclopedia Britannica, Inc., 1968, pp. 357-395.

Simmons, Robert F. Linguistic Analysis of Constructed Responses in Wayne H. Holtzman, Computer-Assisted Instruction, Testing and Guidance, New York: Harper and Row, 1970. pp. 203-221.

Simmons, R.F. Natural Language Question-Answering System. Comm. ACM, 13, 1, (Jan. 1970), pp. 15-30.

Sparck Jones, Karen. Notes on Semantic Discourse Structure. SP-2414, System Development Corporation, Santa Monica, March 3, 1967.

Spolsky, Bernard. Some Problems of Computer-Based Instruction. Behavioral Science, 11, Nov., 1966, pp. 487-496.

Strunk, William, Jr. and White, E.B. The Elements of Style. New York: The Macmillan Company, 1957.

Su, Stanley Y.W. A Computational Model for Paragraph Production. Technical Report No. 71-102, Center for Informatics Research, University of Florida, 1971 (b).

Su, Stanley Y.W. A Semantic Theory Based Upon Interactive Meaning. Computer Science Technical Report #68, University of Wisconsin, 1969.

Su, Stanley Y.W. Managing Semantic Data in An Associative Net. Proceedings of the Symposium on Information Storage and Retrieval, University of Maryland, 1971.

Su, Stanley Y.W. and Harper, K. A Directed Random Paragraph Generator. Reprint No. 13, International Conference on Computational Linguistics, 1969.

U.S. Department of Commerce, National Bureau of Standards. Research and Development in the Computer and Information Sciences, NBS Mongraph 113, Vol. 2, 1970, pp. 26, 92-93.

Vigor, D.B., Urguhart, D. and Wilkinson, A. PROSE-Parsing Recogniser Outputting Sentences in English. Machine Intelligence 4. B. Meltzer and D. Michie (eds.) Edinburgh University Press, 1969. pp. 271-284.

Weizenbaum, Joseph. Contextual Understanding by Computers. Comm. ACM, 10, 8, (August 1967), pp. 474-480.

Weizenbaum, Joseph. ELIZA-A Computer Program for the Study of Natural Language Communication Between Man and Machine. Comm. ACM, 9, 1 (January 1966), pp. 36- 45.

Wilks, Y. Computable Semantic Derivations. SP-3017, System Development Corporation, Santa Monica, January 15, 1968.

Wilks, Y. Interactive Semantic Analysis of English Paragraph. Reprint No. 8, International Conference on Computational Linguistics, 1969.

Winkler, G.P. (ed.) The Associated Press Stylebook. New York: The Associated Press. 1968.

Woods, W.A. Procedural Semantics for a Question-Answering Machine. Proc. AFIPS 1968 Fall Jt. Comput. Conf., Vol. 33, Thompson Book Co., Washington, D.C. pp. 457-471.

Woolley, G.H. Automatic Text Generation. Reprint No. 37, International Conference on Computational Linguistics, 1969.

Yngve, V.E. Random Generation of English Sentences. Proc. 1961 International Conference on Machine Translation of Languages and Applied Language Analysis. Teddington, H.M.S.O., London, 1962, pp. 66-80.

Zinn, Karl L. and McClintock, Susan. A Guide to the Literature on Interactive Uses of Computers for Instruction. A Series One Paper from ERIC at Stanford. Second Edition. 1970.

Zinn, Karl L. Computer Technology for Teaching and Research on Instruction. Review of Educational Research, Vol. 37, 5:618-634, Dec., 1967.